基于结构化地质数据的清洗方法

——以南宁盆地地质数据为例

赵 勇[1,2]，文诗宝[1,2]，卢 鹏[3]，黄梅婷[1,2]

（1. 南宁市勘测设计院集团有限公司，南宁 530022；2. 南宁市浅表地质大数据工程技术研究中心，南宁 530022；
3. 广西路桥工程集团有限公司技术与信息分公司，南宁 530022）

摘要：多源异构的地质大数据在结构化过程中难以保证其准确性。通过构建南宁盆地标准地层体系，采用南宁地质空间大数据分析系统的基于钻孔全自动建模方法创建三维地质模型，并考虑工程地质分区、年代地层、地质成因类型及地层岩性4个因素辅助快速清洗与处理已结构化的入库数据。结果表明，根据三维地质模型出现的不连续区、突变区及拓扑错误区能快速查找结构化数据重复、异常、缺失等问题，有效提升错误识别速度和识别率。

关键词：结构化地质数据；标准地层；地质数据清洗；三维地质模型

中图分类号：TP391 **文献标志码**：A **文章编号**：1671－1807(2025)11－0017－06

随着大数据、数字孪生等技术的快速发展，建设智慧城市已经成为国家发展的重要课题[1]。地质数据作为城市空间基础数据的重要组成部分，可为城市数字化提供工程地质、水文地质、地质灾害等有用数据。地质数据具有可重复开发利用及长期服务社会的特点，合理利用已有的地质数据能提高工作效率、减少资源浪费。然而地质数据挖掘及应用需要满足数据一致性、准确性、完整性等要求[2]，因此对地质数据进行清洗获取标准、干净、连续的数据至关重要。

数据清洗的研究最早出现在美国[3]，随着新一代信息技术发展，数据类型及数据量快速增长，对数据清洗方法要求越来越高。国内学者对数据清洗技术研究可概括为通用方法研究和特定行业的方法研究。在通用方法研究中出现了基于统计学、机器学习等方法。例如，潘婷婷[4]利用哈希算法进行文件内容判断及数据清洗；杨尚林[5]提出层次约减分类清洗和属性约减关联清洗的策略，对多源异构数据进行清洗；孙辞海等[6]提出了基于确定度的交互式数据清洗方法。尽管这些方法通用性强，但主要解决数据重复性和逻辑性的错误。针对特定行业的数据清洗，陈彤[7]考虑石油数据特点，基于Hadoop平台采用聚类分区方法实现数据重复值、异常值、缺失值的清洗；孙乐乐[8]通过分析国土空间大数据特点提出了规则引擎驱动下的数据清洗方法；魏泰等[9]用改进孤立森林算法对风电机组异常数据进行清洗；李立生等[10]针对电网中海量故障数据，运用神经网络和深度学习提出了基于稀疏自编码的故障数据聚类清洗方法。虽然上述数据清洗方法在各自领域都很实用，但用于地质行业针对性不强，难以保证地质数据清洗后的合理性及准确性。

目前，对地质数据清洗的研究较少，研究主要为数据结构化前的重复值、异常值及不完整值[11-12]的处理。陈宇鹏[13]通过改进近邻差值算法对地质灾害数据进行清洗，但仅对监测数据的异常情况进行分析且未对属性数据进行研究。地质数据作为一种时空数据，包含很多对地质现象和地质过程的定性理解、定量估算和关系描述[14-15]；地质数据质量对技术人员的专业程度要求高，结构化地质数据需结合地层沉积规律及地质构造等进行综合判断，常规数据清洗方法很难满足要求。

针对上述问题，以南宁市实际工程勘察地质数据为例，通过构建南宁盆地标准地层体系，采用基于MapGIS研发的南宁地质空间大数据分析系统钻

收稿日期：2024-11-25

基金项目：南宁市优秀青年科技创新创业人才培育项目(RC20220201)；南宁市创新创业领军人才“邕江计划”(2020016)

作者简介：赵勇(1991—)，男，广西桂林人，硕士，工程师，研究方向为三维地质可视化、地质与岩土工程；通信作者文诗宝(1995—)，女，广西玉林人，硕士，工程师，研究方向为地质大数据；卢鹏(1991—)，男，广西柳州人，硕士，工程师，研究方向为岩土工程、测绘等；黄梅婷(1997—)，女，广西都安人，工程师，研究方向为地质大数据。

孔全自动建模功能创建钻孔和地质体三维模型，并分别从工程地质分区、年代地层、地质成因及岩性4个方面对已结构化入库的地质数据进行清洗，旨在为地质数据清洗方法研究提供一些新思路。

1　数据清洗方法

对于地质数据的常规清洗，直接采用南宁市地质空间大数据分析系统能实现钻孔重复性清洗，钻孔原始分层厚度及钻孔深度逻辑性清洗，钻孔原始分层与标准分层存在的矛盾、缺失、重复的清洗。对于已结构化的地质数据，通过构建标准地层体系并将钻孔地层分层数据标准化后建立三维地质模型，以三维地质模型为基础将地下空间可视化，对岩性属性、地层空间关系等进行分析，清洗不符合地质特征的“脏数据”，有效减少地质专业技术人员工作量，确保结构化数据的准确性及有效性。

1.1　标准地层体系构建

标准地层体系的构建不仅能统一南宁盆地地层划分标准，还能快速直观的读取分析地层信息，同时也是实现基于钻孔全自动建模方法的前提。

1.1.1　南宁盆地标准地层划分

根据南宁盆地实际情况，考虑工程分区、年代地层、地质成因、岩性4个因素创建标准地层体系，详见表1。

(1)依据南宁盆地工程地质分区情况，可划分为侵蚀堆积河谷阶地松散土区、剥蚀丘陵陆相碎屑岩区、剥蚀高丘陆相碎屑岩区、溶蚀残峰坡地碳酸盐岩区[16-17]。

(2)根据南宁盆地主要年代地层情况，由新到老分别为第四系、第三系(分新近系和古近系)、白垩系、二叠系、石炭系、泥盆系、寒武系[18]。

(3)参考《工程地质手册》[19]，主要对南宁盆地第四系地层岩土体成因进行划分。

(4)归纳南宁盆地河流阶地及下伏主要的岩土层[20]。

1.1.2　编码规则

为了将标准地层信息转化为机器能识别的编码，同时考虑编码的简洁性和可扩展性，采用8位阿拉伯数字的编码方式[18]。标准地层由4级8位数字组成，第1级为工程分区码、第2级为年代地层码、第3级为地质成因码、第4级为岩性码。编码规则如图1所示，如通过编码11030203能快速识别其代表侵蚀堆积河谷阶地松散土区更新统望高组冲积黏土。

1.2　三维地质模型创建

三维地质建模采用南宁市地质空间大数据分析系统的三维建模子系统，其继承了MapGIS平台在三维地质建模技术上高效、智能的优势[21]。系统提供基于钻孔全自动建模、基于剖面的半自动交互建

表1　标准地层编码表

工程分区	年代地层	地质成因	岩性
侵蚀堆积河谷阶地松散土区(11)	Q_4全新统小河冲洪积、邕江河漫滩冲积层(01)	人工堆积(01)	填土(01)
剥蚀丘陵陆相碎屑岩区(12)	Q_4^g全新统桂平组(一级阶地)(02)	冲积(02)	泥炭、淤泥、淤泥质土(02)
剥蚀高丘陆相碎屑岩区(13)	Q_3^{w2}更新统望高组(二级阶地)(03)	洪积(03)	黏土(03)
溶蚀残峰坡地碳酸盐岩区(14)	Q_2^{b2}更新统白沙组(三级阶地)(04)	残积(04)	粉质黏土(04)
	N新近系(05)	坡积(05)	粉土(05)
	E古近系(06)	湖积(06)	粉砂(06)
	K白垩系(07)	沼泽沉积(07)	细砂(07)
	P二叠系(08)	崩积(08)	中砂(08)
	C石炭系(09)	滑坡堆积(09)	粗砂(09)
	D泥盆系(10)	泥石流堆积(10)	砾砂(10)
	∈寒武系(11)	生物堆积(11)	圆砾、角砾(11)
		化学堆积(12)	卵石、碎石(12)
		溶洞堆积(13)	漂石、块石(13)
		成因不明沉积(14)	红黏土(14)
			含砾土(15)
			混合土(16)
			灰岩(17)
			泥岩(18)
			粉砂岩(19)
			砾岩(20)
			白云岩(21)
			硅质岩(22)
			页岩(23)

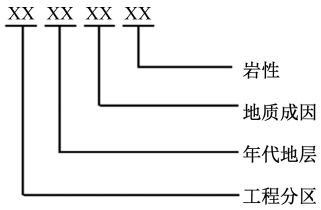

图 1　编码规则

模、基于“分区-拼接”的半自动建模 3 种方法。其中,基于钻孔全自动建模方法具有建模速度快、效率和自动化程度高的特点,很适合通过创建三维地质模型来辅助地质数据的清洗;而其他建模方法虽然建模精度较高,但人工干预过多,建模周期长,不利于用在大规模地质数据的清洗。

基于钻孔全自动建模的原理是从数据库中提取已结构化的钻孔点位和标准地层分层信息,并叠加等值线等约束条件进行插值计算构建地层面模型,最后根据地层之间的接触关系生成实体模型。在插值算法上系统提供了 B 样条、双线性、距离反比 3 种算法,能快速高效地创建符合地质认知的三维地质模型,非常适合数据清洗工作。基于钻孔全自动建模流程(图 2)如下。

(1)对南宁盆地工程勘察钻孔数据进行结构化,并根据标准地层体系编制标准地层编码。

(2)将标准化和结构化的数据导入南宁地质空间大数据分析系统的数据库。

(3)通过三维建模子系统框选研究区钻孔数据、添加约束条件。

(4)选择插值算法(B 样条、双线性、距离反比)。

(5)生成三维地质模型。

2　数据清洗流程

数据清洗可概括为基于二维等值线图的数据清洗、基于三维地质模型的数据清洗和基于地质条件的数据清洗,具体流程如图 3 所示。

(1)按标准地层体系为采集的钻孔地层分层数

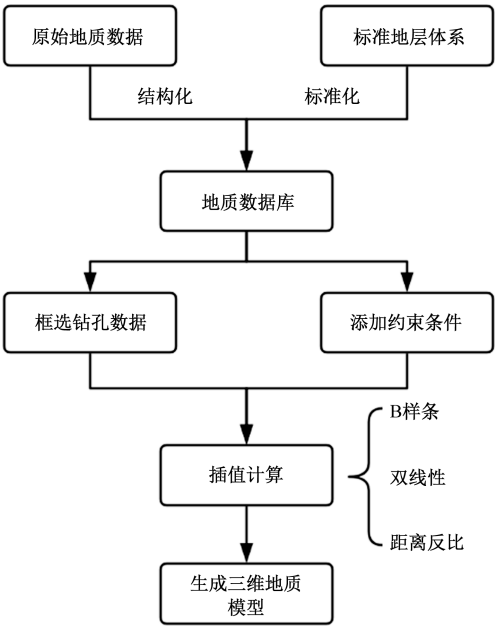

图 2　基于钻孔全自动建模流程

据编制标准地层编号,通过标准地层编号简单直观地了解该地层所属的工程分区、年代地层、地质成因、岩性。

(2)通过南宁地质空间大数据分析系统框选离散数据建立属性二维等值线图,并筛查清洗存在地层厚度异常、不符合地质构造与地层沉积关系的异常点。

(3)框选钻孔数据,用基于钻孔全自动建模功能获得钻孔三维模型及地质体三维模型。通过岩性属性、地层空间关系进行筛选研判,将三维地质模型内拓扑关系、地质构造、地层接触关系等异常问题的地层控制点数据导出,得到属性矛盾、空间关系错误的存疑点。

(4)以区域地形地貌类型、地质成因、沉积年代、岩土体类型及工程地质性质为依据对存疑数据进行清洗、修正,获取准确的地质数据。

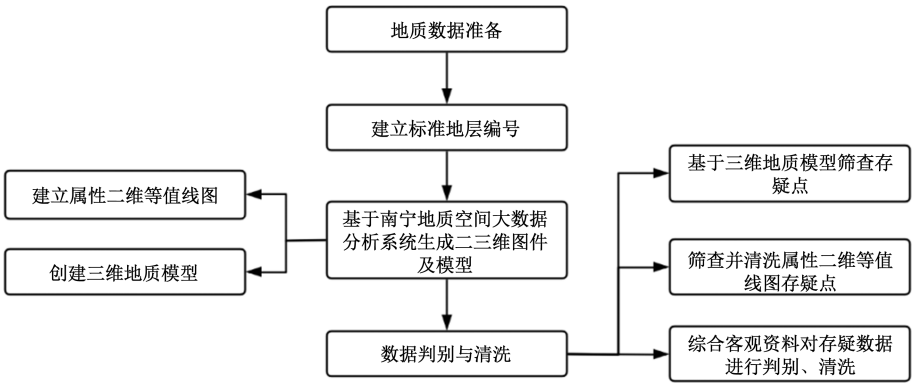

图 3　数据清洗流程

3　示例

3.1　研究区域概况

南宁盆地基底为寒武系砂页岩、泥盆系砂页岩、灰岩及硅质岩、石炭系灰岩和硅质岩等。第三系地层以湖相沉积的泥岩、砂岩及粉砂岩为主，岩性组由老到新分别为瓦窑村组、凤凰山组、古亭组、南湖组、里彩组、北湖组；第四系分布有全新统桂平组、上更新统望高组、中更新统白沙组河流阶地[17]。研究区的位置及工程地质分区情况如图 4 所示。

3.2　数据清洗

3.2.1　数据结构化异常清洗

选取研究区 1 的 696 个钻孔数据创建场地三维地质模型(图 5、图 6)，结果可知在大面积圆砾层之上出现局部第三系泥岩。根据地质资料，该区域上覆地层为第四系更新统望高组冲积层，下伏基岩为第三系泥岩，无断层经过，不存在不整合接触关系。通过简单的区域划分无法判断场地的地层沉积关系及空间分布，而创建三维地质模型能快速直观识别异常区域。经溯源检查原始勘察报告，异常现象是由于数据结构化过程中钻孔坐标识别错误导致的地层空间位置错误。据此，对已结构化数据进行修正，达到清洗的目的。

3.2.2　地层岩性异常清洗

选取研究区 2 的 159 个钻孔数据，创建场地钻孔及地层三维模型(图 7)，对区域内地质拓扑关系、地层沉积关系、地层接触关系异常情况进行检查，判断地质数据的正确性，并对数据进行清洗。

将场地三维地质模型仅显示红黏土、粉质黏土及灰岩(图 8)，模型呈现出外侧粉质黏土包裹红黏土的状态。仅显示红黏土地层时(图 9)，红黏土在三维空间中呈现拉链式、锯齿状，不符合地质规律。经溯源检查，该场地钻孔数据源于两份不同的勘察报告，报告描述的地层岩性情况见表 2，根据上覆地层颜色、液限值及下伏地层岩性可断定报告 1 对岩性的定名不存在问题；而报告 2 钻孔较浅未揭露下伏基岩，但通过三维地质模型推测下伏基岩为石炭系灰岩，且颜色呈红褐色、液限约 40%，符合次生红黏土的特点，又或者是红黏土但因取样试验等原因导则液限偏低，需要进一步验证。

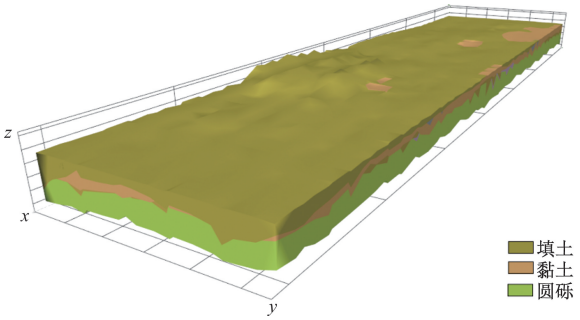

图 5　研究区 1 三维地质模型

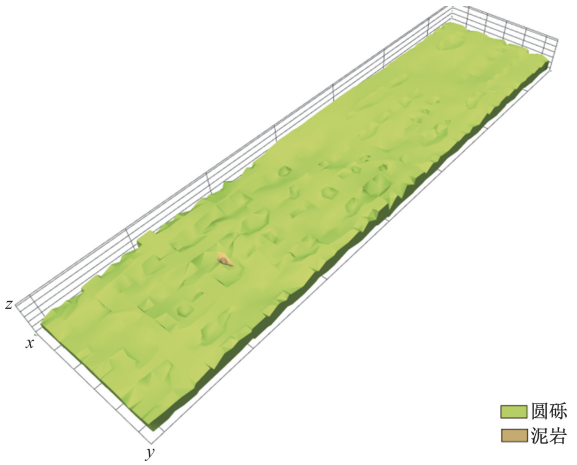

图 6　异常区域三维地层

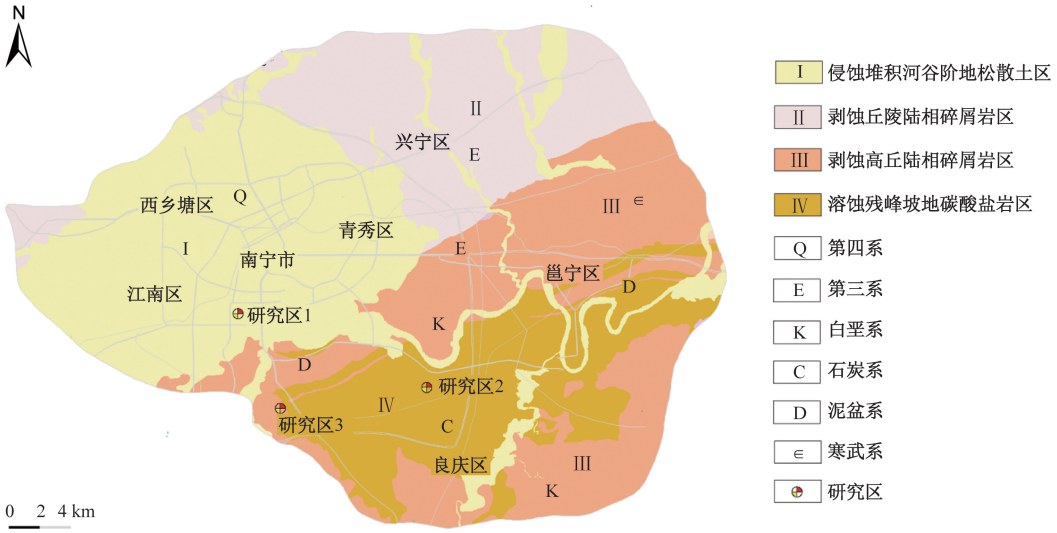

图 4　南宁盆地工程地质分区及研究区位置

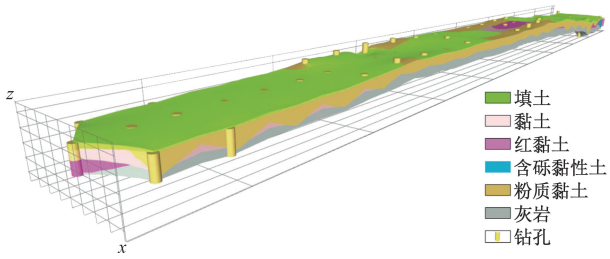

图 7　研究区 2 钻孔及地层三维模型

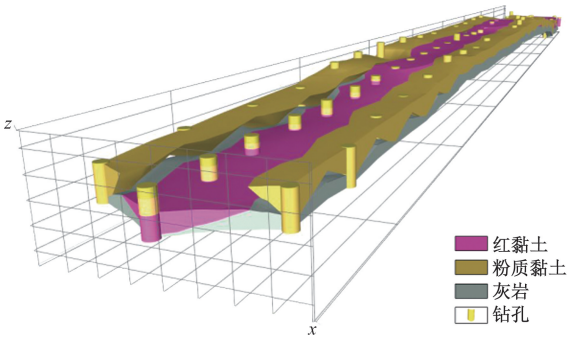

图 8　场地部分地层三维地质模型

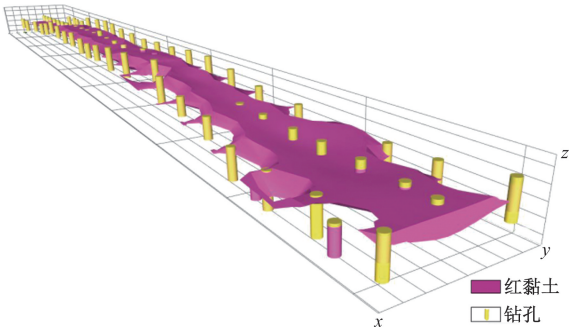

图 9　红黏土三维模型

表 2　研究区 2 岩性描述

位置	类别	红黏土(报告 1)	粉质黏土(报告 2)
上覆地层	颜色	砖红色,局部棕黄色	黄色,褐黄,褐红色,局部灰色
	可塑性	硬塑	硬塑
	干强度	高	中等
	韧性	高	中等
	液限	56.5%～90.9%	30.2%～42.5%
	摇震反应	无	无
	成因类型	残积	残积
下伏基岩	岩性	石炭系灰岩	未揭露

3.2.3　年代地层异常清洗

在研究区 3 选取两份勘察报告共计 255 个钻孔创建三维地质模型(图 10),研究区钻孔分布如图 11 所示,两报告对应地层及其岩性描述见表 3。由于两份报告对砂岩所属的年代地层判别不一致导致创建的砂岩三维模型出现局部空洞(图 12),存在拓扑错误。经核查,该区域钻孔处于工程地质分区泥盆系与白垩系角度不整合接触附近,无断层经过,未

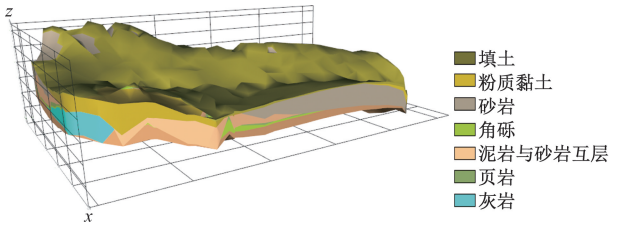

图 10　研究区 3 三维地质模型

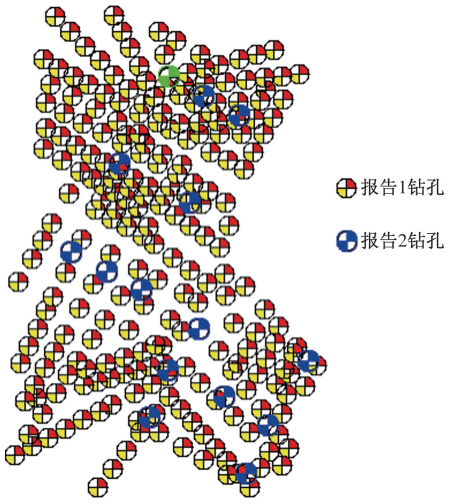

图 11　研究区钻孔分部图

表 3　研究区 3 岩性描述

类别	泥盆系砂岩(报告 1)	第三系砂岩(报告 2)
颜色	紫红、浅紫、褐红、灰黄色	灰色
风化程度	强风化	强风化
结构	粉砂质结构	粉砂质结构
构造	薄层状构造	薄-中层状构造
岩芯	粉状,均为返水捞取,呈角砾、砾砂状	土状,局部呈砂状,节理裂隙发育但多闭合
成因类型	成因不明	湖积

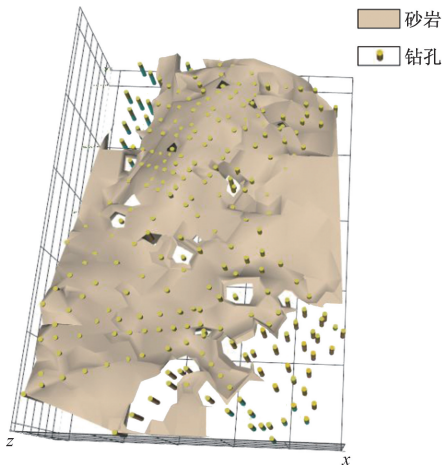

图 12　砂岩三维模型

见出露第三系地层。同时,报告 1 钻孔深度较大,钻孔揭露扁豆状灰岩,为泥盆系五指山组典型地层,

综合判断为报告2的年代地层判断错误。

4 结论

通过构建南宁盆地标准地层体系，并运用基于MapGIS平台研发的南宁市地质空间大数据分析系统创建三维地质模型对已结构化的地质数据进行判别和清洗，得到如下结论。

(1)构建的南宁标准地层体系，既结合了南宁盆地实际地质情况，又考虑了工程地质分区、年代地层、地质成因及岩性因素，不仅能为南宁盆地地层的标准化提供参考，还是基于MapGIS平台三维地质建模及数据结构化的基础。

(2)将钻孔全自动建模速度快、效率和自动化程度高的优势用于地质数据清洗，根据三维地质模型中不连续区、突变区及拓扑错误区，能快速发现地质数据结构化过程中的常规错误、年代地层与岩性判别错误、语义与逻辑错误、地质规律和特征错误，有效提升了错误识别速度和识别率，为地质数据清洗方法研究提供新思路。

参考文献

[1] 薛乾明. 大数据背景下智慧城市空间规划与建设方法[J]. 科技和产业，2023，23(19)：128-135.

[2] 刘军旗，刘强，刘千慧，等. 大数据时代地质灾害数据管理及应用模式探讨[J]. 地质科技通报，2021，40(6)：276-282，292.

[3] 郝爽，李国良，冯建华，等. 结构化数据清洗技术综述[J]. 清华大学学报(自然科学版)，2018，58(12)：1037-1050.

[4] 潘婷婷. 地质空间大数据知识发现与信息提取关键技术研究[D]. 北京：中国地质大学(北京)，2018.

[5] 杨尚林. 基于机器学习的多源异构大数据清洗技术研究[D]. 南宁：广西大学，2017.

[6] 孙辞海，王洪亚，郭开彦，等. 一种基于确定度的交互式迭代数据清洗方法[J]. 智能计算机与应用，2023，13(8)：1-10.

[7] 陈彤. 多源异构海量石油数据的数据清洗技术研究[D]. 青岛：中国石油大学(华东)，2017.

[8] 孙乐乐. 规则引擎驱动下的国土空间大数据清洗方法研究[D]. 昆明：云南师范大学，2018.

[9] 魏泰，贺少雄，胡子武，等. 基于改进孤立森林算法的风电机组异常数据清洗[J]. 科学技术与工程，2024，24(9)：3691-3699.

[10] 李立生，刘洋，卢文华，等. 基于稀疏自编码的故障数据聚类清洗方法[J]. 科学技术与工程，2021，21(15)：6330-6336.

[11] 孙海雪. 地质大数据发现与文本信息分析[D]. 北京：中国地质大学(北京)，2018.

[12] 李婧. 地质大数据发现与信息提取关键技术研究[D]. 北京：中国地质大学(北京)，2016.

[13] 陈宇鹏. 地质灾害数据清洗与修复模型应用研究[D]. 西安：西安工业大学，2024.

[14] 徐德馨，彭汉发，肖杰，等. 城市全空间三维模型数据一体化集成管理关键技术及应用[J]. 地质科技通报，2023，42(1)：388-397.

[15] 田宜平，吴冲龙，翁正平，等. 地质大数据可视化关键技术探讨[J]. 地质科技通报，2020，39(4)：29-36.

[16] 龙睿. 南宁市市区地下空间开发利用适应性评价研究[D]. 桂林：桂林理工大学，2020.

[17] 广西壮族自治区地质矿产局. 南宁市地质系列图集[M]. 南宁：广西壮族自治区地质矿产局，1988.

[18] 陆海丽. 基于南宁盆地工程地质数据库的分区岩土性质及承载力研究[D]. 南宁：广西大学，2016.

[19] 《工程地质手册》编委会. 工程地质手册[M]. 5版. 北京：中国建筑工业出版社，2018.

[20] 张世荣，韦海鑫，刘毛毛，等. 南宁市主要岩土层物理力学参数取值研究[J]. 广西大学学报(自然科学版)，2017，42(4)：1384-1391.

[21] 徐晓雅，王章琼，李雷烈，等. 山岭隧道地上地下一体化三维建模方法[J]. 科学技术与工程，2024，24(8)：3373-3380.

Structured Geological Data Cleaning Method: Taking Geological Data in the Nanning Basin as an Example

ZHAO Yong[1,2], WEN Shibao[1,2], LU Peng[3], HUANG Meiting[1,2]

(1. Nanning Survey and Design Institute Group Co., Ltd., Nanning 530022, China;
2. Nanning Shallow Geology Big Data Engineering Technology Research Center, Nanning 530022, China;
3. Technology and Information Subsidiary, Guangxi Road and Bridge Engineering Group Co., Ltd., Nanning 530022, China)

Abstract: Multi-source heterogeneous geological big data is difficult to ensure accuracy during the structuring process. By constructing a standard stratigraphic system for Nanning basin and using the automatic borehole modeling method of the Nanning Geological Spatial Big Data Analysis System to create a 3D geological model, four factors were considered—engineering geological zoning, age strata, geological genesis types, and lithology—to assist in the rapid cleaning and processing of structured database entries. The results show that discontinuous areas, abrupt change areas, and topological error areas in the 3D geological model can quickly identify issues such as duplicate, abnormal, and missing structured data, effectively improving the speed and accuracy of error identification.

Keywords: structured geological data; standard stratigraphy; geological data cleaning; 3D geological model